Valency in the Prague Dependency Treebank: Building the Valency Lexicon[1]

Markéta Lopatková

Abstract

In this article we focus on valency, which belongs to the core phenomena being captured in the underlying level of the Prague Dependency Treebank (PDT). We present a summary of the basic principles of the applied theoretical framework including proposals for suitable refinement relevant to NLP. The current status of description of valency behavior of verbs, nouns and adjectives is outlined. We present two branches of manual creation of a valency lexicon: (i) the PDT-VALLEX created during the annotation of the PDT and used primarily to obtain consistent annotation, and (ii) the Complex Valency Lexicon VALLEX, where the whole verbal lexemes are processed and other syntactically relevant information is assigned to particular valency frames.

1. Motivation

The Prague Dependency Treebank (PDT) meets the wide-spread aspirations of building corpora with rich annotation schemes. The annotation on the underlying (tectogrammatical) level of language description (Hajičová et al., 2000) – serving among other things for training stochastic processes – allows to acquire a considerable amount of data for rule-based approaches in computational linguistics (and, of course, for 'traditional' linguistics). And valency belongs undoubtedly to the core of all rule-based methods.

PDT is based on Functional Generative Description of Czech (FGD) (Sgall et al., 1986) where the theory of valency has been studied since the seventies. Valency requirements are considered for autosemantic words – verbs, nouns and adjectives (for the references see below). Now, its principles are applied to a huge amount of data – that means a great opportunity to verify the functional criteria set up and the necessity to expand the 'center', 'core' of the language being described.

Within the massive manual annotation, the problem of consistency of assigning the valency structure increases. This was another important impulse, which has led to the decision to create a valency lexicon of verbs, nouns (the theoretical aspects and methodology are refined now) and also adjectives (future plans).

The idea is to create a lexicon containing as much of syntactic-semantic information useful for natural language processing (NLP) as possible.

2. Syntactic vs. semantic approach: an overview of existing projects

In principal, there are two general approaches to the description of valency – a primarily syntactically-based and a primarily semantically-based approach.

[1] Parts of the article (esp. those concerning the Complex Valency Lexicon of Verbs, Section 5.2.) are based on the technical report Lopatková, M., Žabokrtský, Z., Skwarska, K., Benešová, V.: Tektogramaticky anotovaný valenční slovník českých sloves (CKL/UFAL TR-2002-15).

2.1. Description of valency for English

For English, which has the best processed resources, the following projects are the most interesting: FrameNet, LCS Database, PropBank project and Levin verb classes.

2.1.1. FrameNet

The principal goal of the FrameNet project (Fillmore, 2002) is to create a rich lexicon for NLP focusing mainly on verbs and so called 'frame-bearing nouns'.

The FrameNet groups lexical units (pairings of words and senses) into sets according to whether they permit parallel semantic descriptions (i.e. *to tell, to say, to notify* and *to inform*, or their respective meanings, belong with many others to the one semantic class 'Communication'). The verbs from a particular set share the single structure and collection of frame-relevant semantic roles; this frame characterizes the particular meaning of the verbs. The collection of general semantic roles is replaced with the frame specific roles.

Causation

Cause		Affected	Effect
The win caused	*the tree*	*to sway.*	

Communication

Speaker	Message	Addressee	Topic	Medium
Pat	*communicates*	*with Kim*	*about the festival.*	
Pat	*communicates*	*with Kim*		*by the letters.*
Pat	*communicates the message to me.*			

Reciprocality

Protagonists	Prot-1	Prot-2	
	Pat	*fought*	*with Kim.*
Pat and Kim	*fought.*		

Different meanings of a verb can belong to the different groups (e.g. to argue belongs to the 'Quarreling' as well as 'Reasoning' frame.)

For particular frames, it is determined in which way modifications can satisfy semantic and syntactic combinatory restrictions of the respective word (e.g. *to tell, to inform* and *to* notify in their respective meanings can express Addressee as a direct object of the verb).

The frames can create hierarchies where the more specific frames inherit some properties from the more general ones (e.g. frame elements from the 'Quarreling' inherits some properties from 'Conversation' and some properties from 'Disagreeing').

2.1.2. Lexical Conceptual Structure (LCS)

The LCS database (Dorr, 2001) was designed as a semantic representation of predicates and propositions. It describes the semantics of verbs as a combination of semantic structure and semantic content – semantic structure is characteristic for all verbs from one semantic group whereas particular verbs can differ in their semantic content. The lexical item is an oriented rooted graph that bears information on its subject, its objects (arguments) and its 'modificators' and on their obligatoriness / optionality. In addition, their thematic roles are stated as well as restrictions on conceptual categories (also called conceptual POS, as e.g. 'thing', 'event', 'state', 'place', 'purpose', 'manner', 'time').

LCS distinguishes logical arguments (ag, exp, th, src, goal, info, perc, loc, poss, time, prop) and logic modifications (mod-poss, ben, instr, purp, mod-loc, manner, mod-prop) marked with mnemonic labels.

verb *cut down*
lexical item: (act_on loc (* thing 1) (* thing 2)
((* [on] 23) loc (*head*) (thing 24))
(cut+ingly 26)
(down+/m))[2]

cut down: _ag_th,mod-loc(on)

sentence *United States cut down (the) quota.*
(act_on loc (us+) (quota+)
((* [on] 23) loc (*head*) (thing 24))
(cut+ingly 26)
(down+/m))

The LCS is presented as a strictly semantically-based approach with an ambition to be language independent.

2.1.3. Levin Verb Classes

According to the hypothesis stated in (Levin, 1993), syntactic features of verbs are semantically determined and thus syntactic behavior of verbs can lead to their semantic classification. Levin describes syntactic behavior of verbs with respect to possible syntactic alternations and semantic classes are constructed from verbs that undergo a certain number of alternations.

An alternation means a change in the realization of the argument structure of a verb, e.g. 'conative alternation', *Edith cuts the bread* → *Edith cuts at the bread*, or 'middle alternation', → *The bread cuts easily*.

Levin uses the terms 'argument structure' and 'subject' and 'object'; she does not investigate their semantic roles.

This classification, which is very interesting from a theoretical point of view, covers (at least for the time being) only selected meanings of verbs.

2.1.4. PropBank

The main goal of the Proposition Bank project (Kingsbury, Palmer, 2002) is to add semantic annotation to the Penn Treebank. Today, only predicates are processed – an argument structure is assigned to each verb, consisting of arguments (marked Arg0 – Arg5) and modifications (ArgM), with only a minimal specification of the connections between the argument types and semantic roles (Palmer et al, 2001):

He was drawing diagrams and sketches for his patron.
Arg0: he
Rel: drawing
Arg1: diagrams and sketches
Arg2-for: his patron

He keeps st in the fridge.
Arg0: he
Rel: keeps
Arg1: st
Arg2: in the fridge
(see also (Hajičová, Kučerová, 2002))

[2] 'act_on' is a primitive in semantic field 'location'; subject is the thing (type) with the thematic role 'agent' (=1); the only argument is the thing with the thematic role 'theme' (=2); 23='mod-loc' means location with preposition 'on'; 24='mod-loc' is location not required by the verb; the last two nodes specify the manner of the 'location act_on', i.e. *cutting in a downward manner*.

In addition to the annotation, also a valency lexicon of English verbs ('Frame Files') is created. This lexicon stores all the meanings of verbs with their description and examples.

2.2. Czech electronic lexicons of verbs

For Czech there exists a number of valency lexicons, but either their coverage is limited and their forms exclude automatic processing (**'Slovesa pro praxi'** (Svozilová et al., 1997) describing 767 most frequent Czech verbs), or their reliability is not satisfactory, which is caused by automatic processing (**Czech Syntactic Lexicon**, (Skoumalová, 2001)) or they do not store the underlying structure (**BRIEF**, (Pala, Ševeček, 1997) – for each verb it contains all possible combinations of morphemic forms of its complementations). Another problem is connected with the specification of particular meanings of verbs – this problem is not satisfactorily resolved in any lexicon (with the only exception of the lexicon 'Slovesa pro praxi', which uses primarily semantic criteria).

Nevertheless, these lexicons and their underlying methodologies serve as valuable resources for a lexicon satisfying all requirements of complexity, coverage, systematic and consistent treatment of particular phenomena as well as requirements of linguistic adequacy. The lexicon will reflect the both mentioned approaches to the description of valency, the syntactically-based approach as well as the semantically-based one.

3. Theoretical background

3.1. Theory of valency in Functional Generative Description: valency of verbs

Valency theory is a substantial part of the Functional Generative Description (Sgall et al., 1986), a dependency oriented description that serves as our theoretical framework. Valency of verbs has been intensively studied since the seventies (for a comprehensive account, see (Panevová, 1994). The concept of valency primarily pertains to the level of underlying representation of a sentence (i.e. the level of linguistic meaning, called also tectogrammatical level). For NLP, also morphemic representation of particular members of a valency frame is important.

The FGD has adopted a 'middle course' – both syntactic and semantic criteria are used: the first and the second participant is based on syntax behavior of complementations, other inner participants as well as free modifications are detected in accordance with semantic considerations (see below).

The **lexical entry** for a verb enumerates its valency frame(s), i.e. at least one but usually more frame(s) for a verb. The valency frame of a verb (in a broad sense) is interpreted as a range of syntactic elements (verbal complementations) either required or grammatically permitted by this verb. It describes a verb in its primary as well as secondary, 'shifted' use (e.g. *tlačit vůz* [to push a car] vs. *tlačit na někoho* [to urge sb / to press on sb]).

The **verbal valency frame** (in a narrow, strict sense) of particular verb consists of valency complementations (valency 'slots') – inner participants, (arguments, 'actants' in Czech terminology), both obligatory and optional, and obligatory free modifications (adverbial modifications, adjuncts, see below).

On the level of underlying representation, five **inner participants** and a wide scale of modifications are distinguished. The inner participants satisfy the following two conditions:

(i) The combination of participants is characteristic for a particular verb.

(ii) Each participant can appear only once as a complementation of particular verb (if coordination and apposition are not taken into account, not being understood as kinds of dependency).

The participants distinguished here are Actor (or Actor/Bearer, ACT), Patient (PAT), Addressee (ADDR), Origin (ORIG) and Effect (EFF).

Matka.ACT *předělala dětem*.ADDR *loutku*.PAT *z Kašpárka*.ORIG *na čerta*.EFF.
(Panevová)
[Mother re-made a puppet for children from a Punch to an imp.]

Venku prší.
[It is raining.]

On the contrary, **free modifications** (e.g. local, temporal, manner, causal, etc.) can modify any verb and they can repeat with the same verb (the respective constraints are based semantically, rather than syntactically). In PDT about 45 free modifications are distinguished, for the list see e.g. (Hajičová et al., 2002).

V Praze.LOC *se sejdeme na Hlavním nádraží*.LOC *u pokladen*.LOC. (Panevová)
[In Prague we will meet at Main Station near a booking-office.]

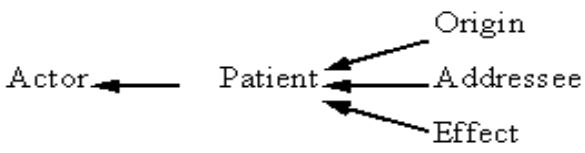
Kvůli dešti.CAUS *musel čekat pod střechou, protože neměl deštník*.CAUS.
(Panevová)
[He had to wait under the roof due to rain because he didn`t have an umbrella.]

The complementations of a verb can be either **obligatory** (i.e. necessarily present at the level of underlying representation) or **optional**. Panevová in (Panevová, 1974-75) stated a **dialogue test** as a criterion for the obligatoriness of inner participants and free modifications.[3] Free modifications are prototypically optional and belong only to a 'valency frame' in a broader sense, but also obligatory free modifications exist (e.g. obligatory Manner in *jednat (s někým*.PAT*) špatně*.MANN [to ill-treat somebody]). It is necessary to strictly distinguish between obligatory complementations, which are not morphematically realized (see Panevová, Řezníčková, 2001), and optional complementations. The relevant functional criteria need further refinement (e.g. in the issue of optionality vs. an obligatory (perhaps general) Addressee with some 'verba dicendi' (verbs of saying)).

	obligatory	optional
inner participants		
free modifications		

Fig.1: Verbal valency frame in strict sense (grey)

The FGD has adopted the concept of **shifting of 'cognitive roles'** in the language patterning (Panevová, 1974-75) see Fig.2. Syntactic criteria are used for the identification of Actor and Patient (following the approach of (Tesniere, 1959)), Actor is the first participant, the second is always the Patient. Other inner participants as well as adverbial modificators are identified on the basis of their semantic roles (as in (Fillmore, 1968), (Fillmore, 1977) and e.g. (Daneš, Hlavsa, 1987) for Czech).



[3] Some of the obligatory participants may be omitted in the surface (morphemic) realization of a sentence, e.g., Actor can be omitted in every Czech sentence. Similarly, also obligatory free modifications are omittable in the surface realization (as e.g. direction for *přijít* [to come], which always means *přijít někam* [to came somewhere]). For the smoothness of the dialogue both speaker and listener must know the necessary information (e.g. from the preceding dialogue or from pragmatics).

Fig.2: Shifting of 'cognitive roles'

In other words – if particular verb has a single inner participant, it is Actor, a verb with two inner participants has Actor and Patient (regardless the semantics). The semantics is taken into account with the third and further participants.

Škola.ACT *začala*.
[The school lesson began in time.]

Bavlně.PAT *se nic*.ACT nevyrovná.
[Nothing is as good as cotton.]

Chlapec.ACT *vyrostl v muže*.PAT. (Panevová)
[A boy grew up to a man.]

Z vašich slov.PAT *plyne*, *že zítra nepřijdete*.ACT.
[It implies from your words that you will not come tomorrow.]

The principle of shifting of 'cognitive roles' can be interpreted as a 'middle course' between the strictly syntactically based and the strictly semantically based approaches.

Verbal complementations can be realized either by single words (esp. by nouns and pronouns in specific cases, but also by e.g. adjectives or adverbs). Or by groups of words – nominal or prepositional groups and coordinated sentence members. Further they can be realized by verbs in infinitive form or by subordinated clauses (with subordinating conjunctions, relative pronouns and adverbs).

For a particular verb, its inner participants have a (usually unique) **morphemic form**, which must be stored in a lexicon (though a prototypical expression of each inner participant exists, as Nominative case for Actor and Accusative case for Patient in active sentence, or Dative for Addressee). Free modifications typically have different morphemic forms connected with their semantics. For example, a prepositional group *na* [on] + Accusative case typically expresses Direction, prepositional group *v* [in] + Locative case usually has local meaning – Where.

The concept **of omissible valency complementations** is reopened with respect to the task of the lexicon. In principle, conditions of omissibility of particular valency slots in the surface realization are not yet clear. Any valency item is presupposeed to be deletable (at least in the specific contexts as e.g. in a question-answer pair). On the other hand, some combinatorical restrictions are probably relevant (see also (Straňáková, 2001)):

předělat [to re-make] ... ACT(1) (ADDR(3)) PAT(4) ORIG(z+2) EFF(na+4)[4]

Matka.ACT *předělala dětem*.ADDR *loutku*.PAT *z kašpárka*.ORIG *na čerta*.EFF.
[Mother re-made a puppet for children from a Punch to an imp.]

The verb *předělat* [to remake] has obligatory Actor, Patient, Origin, Effect and optional Addressee in its valency frame. The variants with omitted Patient and realized Addressee or Effect are not correct Czech sentences:

**Matka*.ACT *předělala dětem*.ADDR *na čerta*.EFF.
[*Mother re-made for children to an imp.]

**Matka*.ACT *předělala dětem*.ADDR.
[*Mother re-made for children.]

[4] We adopt here the notation of Panevová: indices mark particular meanings of the verb; valency members without brackets are obligatory, valency members in brackets are optional and possible morphemic form(s) follow(s) the name of complementation (in brackets; variants are separated by slash '/').

Matka.ACT *předělala loutku*.PAT.
[Mother re-made a puppet.]

Examples of verbal valency frames (only selected meanings of particular verbs are mentioned):[5]

chodit$_1$ [to go / to walk / to pass] ... ACT(1) (PAT(4))

Petr.ACT *chodí do školy.DIR3 pěšky. / Petr*.ACT *chodí dlouhé pochody*.PAT.
[Peter walks to school (= on foot). / Peter goes for long trips.]

chodit$_2$ [to attend] ... ACT(1) DIR3

Petr chodí na gymnázium.
[Peter attends a grammar school.]

chodit$_3$ [to fetch / to go on st] ... ACT(1) INTT(na+4/inf)

Marie.ACT *chodí na borůvky/na nákup/nakupovat*.INTT.
[Mary fetches blueberries. / Mary goes on a shopping.]

chodit$_4$ [to walk out (with sb)] ... ACT(1) PAT(s+7)

Petr.ACT *chodí s Marií*.PAT. (idiom)
[Peter walks out with Mary.]

čekat$_1$ [to wait] ... ACT(1) PAT(na+4)

Rodiče.ACT *čekají na dítě.*PAT *před školou.LOC*.
[The parents wait for they child in front of the school.]

čekat$_2$ [to expect] ... ACT(1) PAT(4/že) (ORIG(od+2))

Petr.ACT *čekal od Jirky*.ORIG *omluvu/že přijde*.PAT.
[Peter expects George`s appology (= an apology from George). /
Peter expects his coming (= that he comes).]

čekat$_3$ [to delay / to trust] ... ACT(1) PAT(s+7)

Věřitel.ACT *jim.BEN čeká s dluhem*.PAT. (idiom)
[The creditor trusts them with a debt.]

čekat$_4$ [to be pregnant] ... ACT(1) PAT(4)

Marie.ACT *čeká s Petrem.ACMP dítě*.PAT. (idiom)
[Mary is with a child (and Peter is its father).]

hovořit$_1$ [to discus] ... ACT(1) ADDR(s+7) PAT(o+6)

Petr.ACT *o svých problémech.*PAT *hovořil s přítelem*.ADDR.
[Peter discussed his problems with his friend.]

hovořit$_2$ [to talk to sb] ... ACT(1) PAT(k+3/na+4)

Otec.ACT *na děti*.PAT *laskavě hovořil.*
[Father talked to his children kindly.]

hovořit$_3$ [to speak (on, upon st)] ... ACT(1) (ADDR(k+3)) PAT(o+4)

Petr.ACT *o své práci*.PAT *hovořil k publiku*.ADDR.
[Peter spoke upon his work in public (= to a public).]

[5] Optional free modifications (not belonging to the valency frame in a strict sense) are in italics in the sentences.

informovat [to inform] ... ACT(1) ADDR(4) PAT(o+4)

Petr.ACT *informoval rodiče*.ADDR *o svém návratu*.PAT.
[Peter informed his parents of his return.]

informovat se [to inform oneself] ... ACT(1) PAT(o+6/na+4)

Petr.ACT *se o jejich práci/na jejich práci*.PAT *informoval*.
[Peter informed himself upon their work.]

jednat$_1$ [to act] ... ACT(1)

Petr.ACT *jednal* (=konat) *rychle.MANN*.
[Pater acted quickly.]

jednat$_2$ [to discuss] ... ACT(1) ADDR(s+7) PAT(o+6)

Petr.ACT *s nimi*.ADDR *jedná* (=vyjednává) *o investicích*.PAT.
[Peter discusses the investments with them.]

jednat$_3$ [to treat] ... ACT((1) PAT(s+7) MANN

Učitel.ACT *jedná* (=zachází) *se žáky*.PAT *špatně*.MANN.
[The teacher mistreats his pupils.]

odpovídat$_1$ [to answer] ... ACT(1) ADDR(3) (PAT(na+4)) EFF(4/že)

Petr.ACT *jim*.ADDR *na dotaz*.PAT *odpovídal vždy pravdu/že* ... EFF.
[Peter always truly answered their question (= he answered them to their question, he a. a truth).
/ Peter answered (= answered them to their question) that ...]

odpovídat$_2$ [to correspond] ... ACT(1) PAT(3)

Řešení.ACT *odpovídá* (=je ve shodě) *požadavkům*.PAT.
[The solution corresponds to the requirements.]

odpovídat$_3$ [to be responsible] ... ACT(1) (ADDR(3)) PAT(za+4)

Rodiče.ACT *odpovídají* (=mají odpovědnost) *za své děti*.PAT.
[Parents are responsible for their children.]

říkat$_1$ [to tell / to speak] ... ACT(1) (ADDR(3)) (PAT(o+6)) EFF(4/že)

Petr.ACT *mu*.ADDR *říkal o Marii*.PAT *pravdu/že je chytrá*.EFF.
[Peter told him the truth about Mary. / Peter told him about Mary that she is clever.)]

říkat$_2$ [to tell / to inform] ... ACT(1) ADDR(3) PAT(o+6)

Petr.ACT *mu*.ADDR *říkal o katastrofě*.PAT.
[Peter told him about the catastrophe.]

říkat$_3$ [to ask] ... ACT(1) ADDR(3) PAT(o+4)

Petr.ACT *mu*.ADDR *marně říkal* (=požádal) *o pomoc*.PAT.
[Peter asked him vainly for help.]

vyhrát$_1$ [to win / to draw] ... ACT(1) PAT(4) (ORIG(na+6))

Petr.ACT *na něm*.ORIG *vyhrál v kartách.REG pět korun*.PAT.
[Peter won five crowns in cards ('from him').]

vyhrát$_2$ [to win] ... ACT(1) (ADDR(s+7/proti+3/nad+7)) (PAT(4))

Petr.ACT *s ním/proti němu/nad ním*.ADDR *vyhrál zápas*.PAT.
[Peter won the match with him.]

vyměnit [to exchange] ... ACT(1) (ADDR(3)) PAT(4) (EFF(za+4))

Petr.ACT *mu*.ADDR *vyměnil staré časopisy*.PAT *za nové*.EFF.
[Peter exchanged him old magazines for the new ones.]

zahájit [to start] ... ACT(1) PAT(4)

Petr.ACT *zahájil schůzi*.PAT *krátkým projevem.MEANS.*
[Peter started the meeting with a short talk.]

žít$_1$ [to live] ... ACT(1)

Petr.ACT *žije v Praze.LOC.*
[Peter lives in Prague.]

žít$_2$ [to scythe / to mow / to reap] ... ACT(1) PAT(4)

Petr.ACT *žal trávník*.PAT *kosou.MEANS.*
[Peter mowed a lawn with scythe.]

3.2. Enriched valency frames of verbs

The 'standard' valency theory applied within FGD is being enriched for the purposes of automatic processing. In addition to the valency slots constituting a valency frame in strict sense (which do not contain optional free modifications) also quasi-valency and typical complementations are stored in the lexicon.

3.2.1. Quasi-valency complementations

Quasi-valency complementations form a new type of complementations on the boundary between inner participants and free modifications (see also (Panevová, 2003)). They are characteristic for their semantics (as free modifications) but they share also important properties with inner participants ((i)-(iii)):

(i) the morphemic form of quasi-valency complementation is predicted by the governing verb;

(ii) there is a limited list of verbs which can be modified by particular quasi-valency complementation;

(iii) each quasi-valency complementation can appear only once as a complementation of particular verb (if coordination and apposition are not taken into account).

Other properties of quasi-valency complementations are shared with free modifications:

(iv) a quasi-valency complementation has typical semantics;

(v) a quasi-valency complementation does not undergo the shifting mentioned above (Fig.2).

Prototypically, quasi-valency complementations are optional (but also obligatory quasi-valency complementations exist, e.g. *zavadit o něco* [to brush against st]).

Determining the set of quasi-valency complementations requires further subtle linguistic inquiry. Now three 'hot candidates' are being proposed:

obstacle (OBST)

uhodit hlavou o větev.OBST [to bump one's head against a bough]

zavadit o stůl.OBST [to brush against a table]

Typically, obstacle has morphemic form *o* [against] + Accusative (i), it is limited to a group of verbs expressing 'negative contact' (ii), it cannot be repeated (iii) and it has typical semantics (iv).

difference (DIFF)

klesat o 5%.DIFF [to fall by 5 percent]

prodloužit o hodinu.DIFF [to prolong by one hour]

Difference has morphemic form *o* [against] + Accusative (i), it is limited to a group of verbs expressing 'change concerning number, strength, capacity, value, etc' and some other verbs as e.g. *vyhrát o délku*.DIFF [to win by length] (ii), it cannot be repeated (iii) and it has typical semantics (iv).

mediator (MDT)

vzít někoho za ruku.MDT [to take sb by his/her hand]

Mediator has morphemic form *za* [for] + Accusative (i), it is limited to a group of verbs expressing 'taking' (ii), it cannot be repeated (iii) and it has typical semantics (iv).

3.2.2. Typical complementations

With free modifications, the semantic constraints are usually mentioned, but they are not specified. The information on typical modifications allows to retain information on valency from existing (printed) dictionaries, which does not belong to the valency frame in the strict sense. What we call typical modification is a free modification that is optional but commonly used with a verb. In general it is not restricted to verb with a particular meaning, usually such a modification modifies the whole group of verbs with similar meaning.

Some of the typical modifications have prototypical form (e.g. Dative case or prepositional group *pro* [for] + Accusative case for Benefactor), the morphemic forms of other modifications are determined by the typical semantics of the modifying members (e.g. prepositional groups *na* [on] + Locative case and *v* [in] + Locative case typically specify Location).

	obligatory	optional	
inner participants			
quasi-valency			
free modifications		typical	

Fig.3: Enriched valency frame (grey)

3.3. Valency of nouns

The valency theory has been primarily established for verbs, which occupy a central position in the sentence structure. The extension on nouns and adjectives has followed, see esp. (Piťha, 1981) and (Panevová, 2000).

Two groups of nouns are distinguished with different valency characteristics – nouns derived from verbs, **deverbal nouns** and **primary nouns**.

Generally, deverbal nouns 'inherit' in some way the valency frames of their source verbs, however the process of nominalization is quite complex and complicated. The derivation can be accompanied by a reduction of some complementation(s), some complementation can be 'incorporated' etc. The theoretical aspects and methodology are being refined now, see esp. (Řezníčková, 2003).

Describing valency frames of nouns, the set of verbal complementations must be enlarged with special nominal complementations – the inner participants Partitive (MAT, e.g. *skupina lidí*.MAT [group of people]) and Identity (ID, e.g. *město Praha*.ID [city Prague]) and the free modifications Appurtenance

(APP, e.g. *Janův bratr / bratr Jana* [the brother of John]]), Restrictive (RSTR) and Descriptive Adjunct (DES).

Special attention must be paid in the future to verbonominal collocations, i.e. collocations of nouns and verbs that constitute a single lexical unit and as such have valency complementations 'in common'.

3.4. Valency of adjectives

The valency frames of adjectives have been studied by Piťha and Panevová, see (Piťha, 1982) and (Panevová, 1998). They have started from **deverbal adjectives** (i.e. adjectives derived from verbs) which have the same repertoire of inner participants as verbs. A deverbal adjective shares its valency frame with the original verb. One regular difference is present:

(i) One of the expected valency slots is 'absorbed' by the word that is modified by the examined adjective (i.e. by the governor of the adjective, see (Panevová, 1998)).

žít$_2$ [to lead a life] … ACT(1;obl) PAT(4;opt)

Petr.ACT *žil smysluplný život*.PAT.
[Peter led a meaningful life.]

→ *žijící* … PAT(4) ACT absorbed

Petr žijící smysluplný život.PAT
[Peter leading a meaningful life]

→ *žitý* … ACT(7) PAT absorbed

život žitý vnímavým člověkem.ACT
[a life led by sensitive man]

In addition to the same list of inner participants and free modifiers as the verbs have, **non-deverbal adjectives** have also specific modifiers of their comparatives and superlatives. The principle of 'shifting' is not applied here.

The description of the theoretical aspects of valency of adjectives need further refinement.

4. Valency in the Prague Dependency Treebank

The Prague Dependency Treebank has a three-level structure:

(i) full morphological annotation – to each word a (disambiguated) tag specifying complete morphological information is assigned;

(ii) annotation on the analytical level – to each sentence an analytical tree is assigned, i.e. dependency tree describing the surface syntactic structure of a sentence;

(iii) annotation on the underlying (tectogrammatical) level – to each sentence its tectogrammatical tree structure, TGTS is assigned.

The valency belongs to the core concepts of the level of underlying representation. The PDT has completely adopted the conception of 'standard' valency of the Functional Generative Description (Sections 3.1, 3.3. and 3.4.).

Verbs

All autosemantic verbs have assigned their valency frames (in a strict sense) – inner participants (obligatory as well as optional ones) and obligatory free modifications must be specified. If any of the obligatory valency members is not present in the surface realization of a sentence (e.g. general

participant or actual ellipsis), it must be restored in the respective TGTS. For the detailed rules see e.g. (Hajičová et al., 2001) and (Hajičová et al., 2002).

Nouns

In the 'large collection' all obligatory complementations are assigned to deverbal nouns with the suffixes *–ní/-tí* (*zpracování* / *zpracovávání* [processing], *dobytí* [conquering]);[6] if such a complementation is not realized in a surface shape of a sentence (e.g. general participant or actual ellipsis) it must be restored in a TGTS

Concerning other nouns (e.g. *výběr* [choice], *příchod* [arrival], *román* [novel], *skupina* [group]) only complementations realized in the surface shape of a sentence are annotated, obligatory complementations not realized are not restored, see (Hajičová et al., 2002).

In the so called 'model collection' all obligatory participants and free modifications are assigned to all nouns, the elided ones are restored.

Adjectives

In the 'model collection' all obligatory participants and free modifications are assigned to deverbal adjectives, the elided ones are restored.

5. Valency lexicon of Czech verbs

Within the massive manual annotation, the problem of consistency of assigning the valency structure increases. This problem can be (at least to a great extent) solved with the valency lexicon.

For the purposes of annotation on tectogrammatical level a valency lexicon is intensively built up in two branches:

(i) The first branch is represented by the lists of valency frames being created and used by annotators during their work. It contains valency frames of words (verbs and nouns) in their particular meanings (as they appear in the PDT) and serves for consistency of annotation.

(ii) The second branch is represented by the valency lexicon, in which the words (only verbs in this stage) are analyzed in the whole complexity, in all their meanings. Rich syntactic information is assigned to particular valency frames, including e.g. control and reciprocity.

There have been also attempts to create a valency lexicon of verbs automatically, exploiting an annotation on the analytical level of PDT (i.e. the level describing surface syntactic structure of the sentence). The most important attempt is described in (Sarkar, Zeman, 2000).

Here we focus on manually creating the valency lexicon of Czech verbs (as they are best processed – the principles of the valency of nouns are refined now, see (Řezníčková, 2003), the valency lexicon of adjectives belongs to the future plans).

5.1. Lists of valency frames in PDT (PDT-VALLEX)

The annotators construct lists of valency frames during their work. These lists (also called PDT-VALLEX) contain valency frames of verbs and nouns in their particular meanings (as they appear in the PDT); the lexeme as a whole is not analyzed. These lists serve first of all for consistency of annotation – for a particular verb or noun the annotators choose one of the existing valency frames from the list or add the respective valency frame to the list (if the verb has not been used in this meaning in the sentences

[6] More precisely, this is valid for these nouns when expressing an event. In their resultative usage they are treated in the same way as other nouns.

processed). (A set of tools has been developed to ease the searching and appending the lists of valency frames, see (Hajič et al, 2001)).

A valency frame consists of particular valency members, for each valency frame the following information is specified:

(i) 'functor' (the kind of the respective inner participant or free modification),

(ii) the type of complementation (obligatory, optional or typical),

(iii) possible morphemic form(s) of the complementation and

(iv) example(s) of usage.[7]

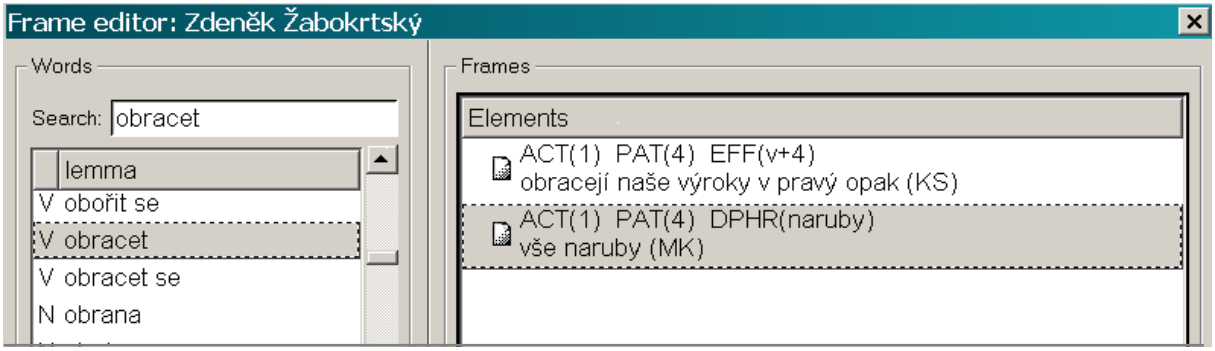


Fig. 4.: Valency frames of the verb *obracet* [to turn] in the list of valency frames in PDT

In these days, the list of verbal valency frames contains about 4 720 verbs with 7 160 valency frames (i.e. over 1,5 frames per verb).[8] Concerning nouns, there are about 1320 lexemes with 1420 valency frames (mostly automatically generated from annotated sentences).

5.2. VALLEX – the Complex Valency Lexicon of verbs

5.2.1. Core information – valency frames

Valency frames

The goal of the Complex Valency Lexicon (VALLEX in the sequel) is to describe the whole verbal lexemes (i.e. verbs in all their meanings).[9]

The lexical entry in the VALLEX is composed of a set of valency frames; one valency frame typically corresponds to one meaning (but not necessarily, see below). A valency frame consists of a sequence of valency members, in each valency frame the following information is specified:

(i) 'functor' (the kind of the respective inner participant or free modification),

(ii) the type of complementation (obligatory, optional or typical) and

(iii) possible morphemic form(s) of the complementation.[10]

---

[7] The valency member is represented by a functor followed by brackets with possible morphemic form(s) of a complementation; the type of complementation is specified with the brackets – round brackets mean obligatory member, squared bracket optional member.
E.g. one of the valency frame of the verb *vyprávět* [to tell] ... ACT(1) ADDR(3) EFF(4,že) PAT[o+6]

[8] These numbers will slightly change as the annotation of PDT continues (changes in tens of verbs mostly).

[9] At least all primary and secondary meanings are described; the problem of a complex description of idiomatic and frozen collocations is still open.

[10] The valency member is represented by a functor followed by brackets with possible morphemic form(s) of a complementation (number means case, preposition+number indicates prepositional group, inf means infinitive and subordinated conjunction represents dependent clause) and type of complementation (separated by ';').
E.g. one of the valency frame of the verb *vyprávět* [to tell] ... ACT(1;obl) ADDR(3;obl) PAT(o+6;opt) EFF(4,že;obl)

The respective meaning(s) of the verb is/are specified by the synonym(s) or a gloss (attribute 'synon') and by example of usage (attribute 'example').

The functors are ordered in accordance with systemic ordering (SO), which reflects unmarked word order (see (Sgall et al., 1986)), with the only exception – all inner participants precede free modifications.

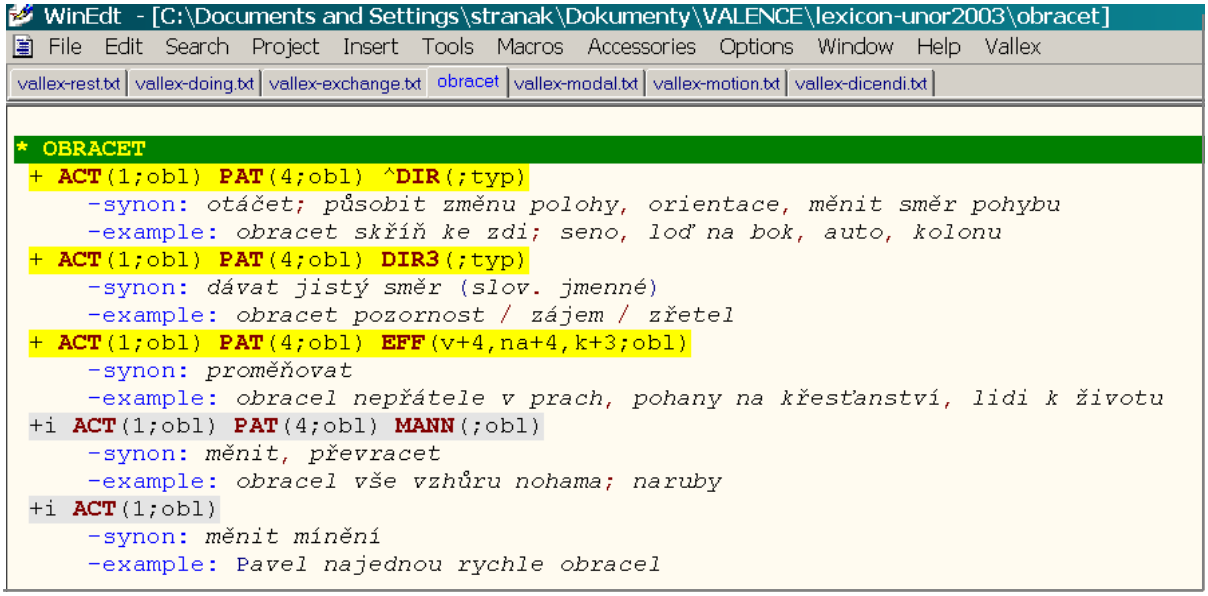


Fig. 5.: Valency frames of the verb *obracet* [to turn] in the VALLEX

The VALLEX draws information from existing Czech dictionaries, namely the BRIEF ((Pala, Ševeček, 1997), for each verb containing all possible combinations of morphemic forms of its complementations) and 'Slovesa pro praxi' ((Svozilová et al., 1997), a semantically-based valency lexicon of most frequent Czech verbs). Also printed lexicons of Czech are consulted (Slovník spisovného jazyka českého (SSJČ), Slovník spisovné češtiny (SSČ), Slovník českých synonym (SČS) and Slovník české frazeologie a idiomatiky (SČFI)).

Setting off particular valency frames

A Czech verb as a whole – verb lexeme – is an abstract unit made up by all meanings of a particular verb. A verb lexeme consists of a set of lexical units, each of which represents a single meaning and has specific syntactic characteristics.

A lexical entry in the VALLEX corresponds to the whole lexeme. Each lexical unit is specified by its valency frame and its meaning – any change either in valency frame or in the meaning leads to a change of lexical unit.

This implies that the division of particular valency frames is based on distinguishing particular meanings of a verb (and on changes in syntactic behavior). Unfortunately, no generally accepted criteria for distinguishing particular meanings of verbs exist – the existing lexicons of Czech verbs differ in specification of particular meanings of verbs (e.g. in BRIEF (Pala, Ševeček, 1997) this problem is not solved at all).[11] The formulation of suitable criteria and their further refinement belongs to the core and most complicated problems of the project.

[11] For each verb lexeme, the BRIEF contains all possible combinations of morphematic forms of its complementations; particular lexical units are not distinguished.

In the VALLEX the following principles are adopted (they are exemplified below):

(i) the difference in the meaning is a necessary but not sufficient condition for setting off two (or more) valency frames – the (slight) difference in the meaning is ignored if more lexical units do not differ syntactically; i.e. the annotators rely rather on syntactic behavior of a verb (including additional syntactic information, see the next Section);

(ii) two different lexical units can have identical valency;

(iii) the change in morphemic realization signalizes the possibility of different meanings; on the other hand

(iv) particular complementation in a valency frame can have morphemic variants (if the meaning is 'sufficiently close').

postavit$_2$ [to raise / to build up] ... ACT(1;obl) ADDR(3;opt) PAT(4;obl) ORIG(z+2;opt)

postavit sochu [to raise a statue]

postavit budovu [to build up a building]

postavit model letadla [to construct a model of a plane]

These three usages of the verb *postavit* are described in one valency frame – the difference in the meaning is not taken into account.

absolvovat$_1$ [to pass / to finish] ... ACT(1;obl) PAT(4;obl)

absolvovat školu [to pass a school]

absolvovat$_2$ [to go through] ... ACT(1;obl) ADDR(4;obl) PAT(s+7;obl)

absolvovat operaci [to go through an operation]

Though *absolvovat*$_1$ and *absolvovat*$_2$ have an identical valency frame the difference in meaning has to be reflected by distinguishing two lexical units.

hlásit se$_2$ [to be counted among sb] ... ACT(1;obl) PAT(k+3;obl)

hlásit se ke komunistům [to be counted among communists]

hlásit se$_4$ [to apply for st] ... ACT(1;obl) PAT(o+4;obl)

hlásit se o svá práva [to apply for oneself rights]

The change in morphemic realization signalizes different meanings and thus two lexical items *hlásit se*$_2$ and *hlásit se*$_4$.

učit$_1$ [to teach] ... ACT(1;obl) ADDR(4;obl) PAT(3,4,inf,že,zda,aby,jak;obl)

Učitel učí žáky matematice/matematiku/pracovat/...
[Teacher teaches his pupils mathematics/to work/...]

This lexical unit has several possibilities how to express the obligatory Patient.

5.2.2. Additional syntactic information for particular valency frames

The VALLEX is built with an ambition to store all syntactic information needed for NLP in one resource. In addition to the core information on valency behavior of verbs – valency frames – it contains also supplementary information associated with particular valency frames.

Reflexivity

Both from theoretical and practical point of view it is necessary to state functional criteria for decision whether a verb with the reflexive pronoun *se/si* constitutes a separate lexeme or belongs to the lexeme without *se/si* (then it is useful to state the function of the reflexive pronoun).[12]

In the VALLEX, the following functions of *se/si* are distinguished (attribute 'refl.'), see also (Králíková, 1981):

(i) *se/si* is a part of verbal lemma for so called 'reflexivum tantum' (the verb does not exist without this particle, i.e. *bát se* [to be afraid]), value 'AuxT';

(ii) *se* is a part of an analytic form of a verb (i.e. reflexive passive), value 'AuxR';

(iii) *se/si* fills one valency slot (it marks an object identical with the Actor, *mýt se* [to wash oneself], *koupit si* [to buy st for oneself]), value 'Objse' or 'Objsi';

(iv) *se/si* is a part of verbal lemma for so called 'derived reflexives' (i.e. unintended *zabít se* [to be killed] and spontaneous *vlny se šíří* [waves defusse], or *vrátit se* [to return]), value 'derived';

(v) reciprocal *se/si* is described in the attribute 'reciprocity' (see below).

Just 'reflexives tantum' and 'derived reflexives' constitute separate verb lexemes.

Reciprocity

Reciprocity means the possibility of a member of a valency frame to enter the symmetric relation with other member of this frame. Following (Panevová, 1999), the reciprocal usage is considered as a special usage of a basic valency frame – it is only necessary for the members entering the relation of reciprocity to be marked (attribute 'reciprocity' in particular frames in the VALLEX).

představovat$_1$ [to introduce (sb to sb)] … ACT(1;obl) ADDR(3;obl) PAT(4;obl) EFF(jako+4;opt)

Spolužáci se představovali (sobě navzájem).
[The school-mates introduced (themselves) one to another.]

reciprocity: ACT-ADDR-PAT

Panevová describes mainly reciprocity among inner participants, albeit she supposes also reciprocity between inner participant and free modifications. Based on the data processed, it is necessary to admit combinations of participants and modifications in a considerably broader scale (the restrictions are semantically based, i.e. both (all) reciprocal members are animal).

Kamarádky si (navzájem) hlídaly děti (když to některá z nich potřebovala).
[The friends in return took care for children (when any of them need it).]

-reciprocity: ACT-BEN

Bratři mluví za sebe (navzájem / jeden ve prospěch druhého).
[The brothers spoke instead of / in favor of the other.]

-reciprocity: ACT-SUBST

Petr a Pavel si (spolu) hrají.
[Peter and Paul play together (one with the other).]

-reciprocity: ACT-ACMP

[12] The methodology of assigning reflexivity is proposed, the annotation will start this year.

Manželé se zařídili podle sebe.
[Husband and wife (a married couple) have adapted one to the other.]

-reciprocity: ACT-NORM

Control

The term 'control' primarily relates to a certain type of predicate (verbs of control) that can have an infinitive complementation (regardless its functor).[13] With such verbs two co-referential expressions are related, a controller and a controlee. Then controlee is a member of a valency frame that would be a subject of infinitive,[14] controller is coindexed member of relevant valency frame of head verb (see (Panevová, 1996)). In such a case, a controller (its functor) is marked in the lexicon.

nabídnout [to propose] … ACT(1;obl) ADDR(3;obl) PAT(4,inf,že,aby,ať;obl) EFF(7,jako+4;opt) RCMP(za+4;typ) AIM(k+3,na+4;typ)

Petr mi nabídl napsat ten dopis místo mě.
[Peter proposed to me that he would write the letter instead of mine.]

-control: ACT

Petr mi nabídl přespat u něj.
[Peter proposed to me that I can sleep in his room.]

-control: ADDR

There are also verbs where the subject of their infinitive complementation is not expressed in their valency frame. Then the controller is marked by 'ex'. This label marks also yet unclear cases with impersonal verb constructions:

patřit se [to beseem] … ACT(1,inf,že,aby;obl)

Patřilo se přijít včas.
[It is beseem to come in time.]

-control: ex

Aspect and aspectual counterparts

In the VALLEX the attribute 'aspect' ('vid') can assume the following values: perfective ('dok'), imperfective ('ned'), both perfective and imperfective ('dokned') and iterative ('nás'). This attribute is asserted for each valency frame.

hodit$_1$ [to throw] … ACT(1;obl) ADDR(3;opt) PAT(4;obl)

-vid: (dok); házet (nedok)

chodit$_2$ [to attend] … ACT(1;obl) DIR3(;obl)

-vid: (ned); chodívat (nás)

orientovat$_1$ [to orient] … ACT(1;obl) PAT(4;obl) LOC(;obl)

-vid: (dokned)

The separate processing of aspectual counterparts was proved to be useful in the initial phases of annotation. Later on the lexicon was restructured and the aspectual counterparts (in a strict sense)[15] have

[13] More generally, verbs of control have two co-indexed complementations – besides an infinitive one valency member can be realized by nominalization.
[14] In some cases the morphemic realization is structurally excluded.
[15] Two verbs are considered as the aspectual counterparts in a strict sense if they have the same meaning and they differ only in the category aspect ('dok' and 'ned'), thus they belong to a single lexeme.

been linked for particular valency frames (and annotation has been corrected or frames added if necessary).[16]

odpovídat$_1$ [to answer] ... ACT(1) ADDR(3) (PAT(na+4)) EFF(4/že)

-vid: (ned); odpovědět (dok)

odpovídat$_2$ [to correspond] ... ACT(1) PAT(3)

-vid: ned

odpovídat$_3$ [to be responsible] ... ACT(1) (ADDR(3)) PAT(za+4)

-vid: ned

Moreover, the verbs have been grouped in so-called 'clusters'. The clusters are typically created by aspectual counterparts (in a strict sense) and by a corresponding iterative verb (if this exists). Eventually, also prefixed verb(s) can be added to the cluster if its (their) meaning(s) is (are) very close to the original verb and its (their) valency behavior does not differ (then such verbs are marked as 'dok1'/'ned1', 'dok2'/'ned2', … with respect to their aspect).

Within the cluster, particular valency frames are listed; for each valency frame, the element(s) of the cluster is (are) specified for which the frame is valid.

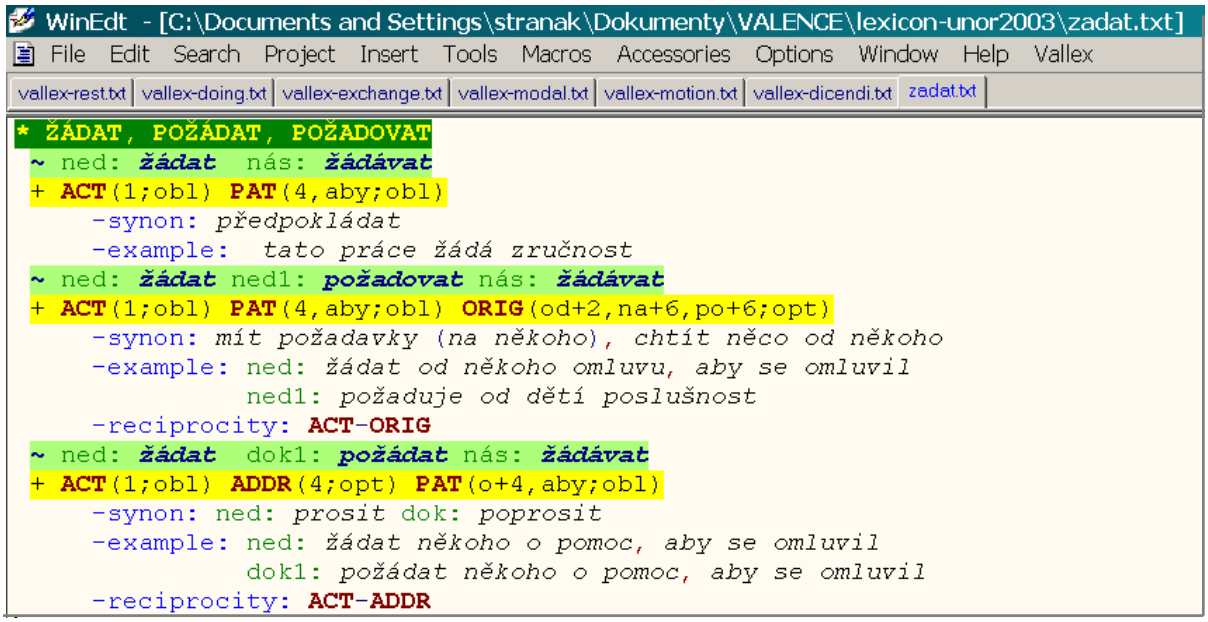


Fig. 5.: Valency cluster in VALLEX – the verbs *žádat* [to ask / to demand], *požádat* [to ask / to request / to apply] and *požadovat* [to demand / to require]

This structure makes it possible to describe reasonably also groups of related verbs with complicated aspectual behavior; consider e.g. the triple *žádat* [to ask / to demand], *požádat* [to ask / to request / to apply], *požadovat* [to demand / to require][17] – these three verbs are grouped in one cluster because they in twos share syntactic-semantic characteristics; see Fig.5.

The clustering just described serves for achievement of consistency in the lexicon and its sustentation. Grouping verbs to clusters can be seen as additional information, which can be filtered out.

[16] This two-fold annotation serves as a good test of agreement among annotators.

[17] These verbs have the same morphological root. According to the lexicon of Czech SSJČ *žádat* ('ned') has the counterpart *požádat* ('dok') and *požádat* ('dok') has the counterpart *požadovat* ('ned'); some linguists do not consider the aspectual pair *požádat* ('dok') – *požadovat* ('ned').

Possible diatheses, passivization

In the VALLEX verbs in the primary diathesis are stored ('active' frames). The possibility of adding information on secondary diatheses was studied (Lopatková et al., 2002). This information must be added for particular frames. The following types of diatheses will be marked in the future:

(i) the possibility of a periphrastic passive (*kniha byla vydána*)

(ii) *mít*+passive participle (*spolužák měl kancelář přidělenou správcem objektu*)

(iii) *dostat*+passive participle (*dostat vyhubováno*)

(iv) The possibility of a reflexive passive is stated in the attribute 'reflexivity'.

Primary / secondary / idiomatic usage

Whereas in theoretical studies verbs have been described mainly in their primary meaning, now the whole verbs (lexemes) in all their meanings are processed – verbs in their primary and secondary meanings as well as verbal idioms and frozen collocations are treated. This is connected with the effort to reach maximal 'coverage' of texts.

One of the lexical units that constitute one verb lexeme is usually considered as a **primary**, basic one (traditionally the first one in written lexicons), e.g. překvapit přítele dárkem [to surprise friend with a gift] (attribute 'use', value 'prim').

Other lexical units are **derived** from the primary one, with a change in its meaning, e.g. překvapit zloděje při krádeži [to take a thief anawares on burglary].

The description of **frozen collocations** is an important task to be taken up in the lexicon. Frozen collocations and idioms are traditionally understood broadly in Czech linguistics. In the VALLEX frozen collocations without syntactic irregularities are described; their frequency in Czech National Corpus (CNC) serves as a working criterion for their taking up.

In the VALLEX (as well as in the PDT), the following principles have been adopted:

(i) 'to analyze syntactically what can be analyzed' using the standard functors;

(ii) the information on frozen collocation is marked in the attribute 'use', value 'idiom';

(iii) the specific functor DPHR is reserved for the dependent parts of collocations with which the complementation is lexically limited to a single word (or to a restricted set of words) and the collocation cannot be syntactically analyzed.

In the VALLEX the valency frames of a particular verb are ordered with respect to type of their usage – primary (attribute 'use', value 'prim') usage(s) is (are) the first, secondary usage(s) (value 'posun') and then idiom(s) follow (value 'idiom') (with respect to their 'frequency', see below).

přijít do školy.DIR1 [to come to school]

-use: prim

přijít nakoupit.INTT [to come for shopping]

-use: prim

přijít na skvělou myšlenku.PAT [to think out an excellent idea]

-use: posun

přijít o hodinky.PAT [to drop watch]

-use: posun

přijít k penězům.PAT [to obtain money]

-use: posun

přijít do jiného stavu.DPHR [to become pregnant]

-use: idiom

Syntactic/semantic class

Processing whole groups of verbs with similar semantic properties is a principle of good promise, esp. what concerns consistency and completeness of the lexicon. (E.g. the groups of verbs which can be marked as 'exchange' and 'motion' verbs and 'verba dicendi' (verbs of saying) have been processed in parallel with good results.)

As it is not possible to simply adapt any of the existing classifications of verbs, the principles for building syntactic-semantic classes of verbs 'bottom-up' are being formulated now.

The classes are constituted from singe lexical units (not whole lexemes) as particular lexical units may belong to different classes.

Up to these days, the attribute 'class' has been intuitively filled in for about 650 frames. Though the classification is just preliminary (the research is in its initial phase), 15 groups have been established, which serve for consistency checking during annotation.

Pointers to Czech EuroWordNet

EuroWordNet. EuroWordNet, EWN is a multilingual lexical database consisting of national WordNets, lexical databases for several European languages including Czech (Pala, Ševeček, 1999). The WordNets are based on so called synsets ('sets of synonyms'), i.e. sets of words that can be replaced in some contexts. These synsets are linked with the 'Inter-Lingual Index', ILI determining equivalent synsets in different languages (via English).

The possibility of establishing links between valency frames and synsets was tested for about 400 verbs. Approximately one half of them was processed both in the VALLEX and Czech WordNet (preliminary version, 2001). The linking brought up a number of problems (the national WordNets are based on English, not primary on the respective language):

(i) missing synset (no synset in Czech WordNet corresponds to some meaning of a Czech verb);

(ii) redundant synset (no meaning of a Czech verb corresponds to the English equivalent);

(iii) confusing specification of synset (subjective and inconsistent specification of particular meanings of verbs in EWN).

The ideal situation in which there is a 1:1 relation between sets of valency frames of a verb and sets of its synsets is rare.

Nevertheless the advantages of even imperfect links between particular valency frames and respective synsets are obvious.

Frequency in sample of CNC

The auxiliary attribute frequency ('freq.') contains information on testing particular valency frames. Each verb processed is tested with respect to a sample from the Czech National Corpus (CNC) – two files, each of 30 randomly chosen sentences with particular verb, are used. This testing (the first phase during the initial assignment of valency frames, the second one during consistency checking, the third is being prepared in these day) has two main goals:

(i) to verify the completeness of the lexicon (whether the lexicon contains all meanings that have appeared in a sample) – this test may show relatively frequent meanings not described in the existing lexicons;

(ii) to verify distinguishing particular valency frames (and particular lexical units); the possibility of assigning a single valency frame to each occurrence of a verb in samples is tested.

In addition, the numbers of occurrences of particular meanings of tested verbs in the sample allow to order valency frames in the VALLEX, which speeds up human searching in the lexicon.

5.2.3. VALLEX – state of art

For the time being (summer 2003), over 2 480 occurrences of valency frames have been treated, grouped in 858 clusters; this equates approximately 1 450 verb lexemes. The verbs were chosen according to their frequency in the CNC and PDT; the auxiliary to be, which requires a special treatment, has been excluded yet.

These numbers for the annotated verbs are not final – there will be small changes due to the inconsistent specification of (derived) reflexives (this shortcoming is just setting up, see above.)

Together with the verb to be, which will be treated in the foreseeable future, the lexicon covers a significant part of verbs in texts from CNC (about 85 percent on 'running text'), further enlargement is supposed.

The VALLEX is planned to be released for research activities in October 2003 (see http://ckl.mff.cuni.cz/?a=activities).

5.3. Comparison of the PDT-VALLEX and the VALLEX

The lists of valency frames in PDT (PDT-VALLEX, Section 5.1.) and the complex valency lexicon VALLEX (Section 5.2.) represent two branches of manual building valency lexicon of Czech verbs.

The PDT-VALLEX is expanded 'extensively', the particular meanings of verbs are added as they appear in annotated sentences in PDT (and thus its 'coverage' improves relatively quickly, which positively affects 'recall' when used in automatic procedures, see (Hajič, Honetschlager, 2003)). The ending annotation of PDT will limit the expansion of these lists (the final number of verbs is estimated at 4 800 verbs). The way of creating lists of valency frames (verbs are not treated as complex units, the annotators add particular meanings of verbs 'as they need') imposes the necessity of thoroughgoing and time-consuming consistency checking in the final stages of the PDT annotation.

On the other hand, the expansion of the complex valency lexicon VALLEX can be characterized by the fact that the whole verbal lexemes are processed, many syntactic relevant information is added. Significant stress is laid on the maximal consistency of all assigned information. This allows for a linguistically adequate representation and description of valency properties of verbs. This approach decidedly impresses 'precision' in automatic procedure see (Hajič, Honetschlager, 2003).

The two lexicons met in a common point: they were merged in December 2001. Since then they have been developed independently. Their final merging is forseen, depending on the completion of PDT annotation and consistency checking of the PDT-VALLEX:

(i) the comparison of the PDT-VALLEX and the VALLEX will verify the criteria adopted and lead to their refinement; the information on particular meanings of verbs treated in both lexicons will be unified;

(ii) the PDT-VALLEX will serve as a valuable source of processed verbal meanings for the VALLEX (verbs not treated in the VALLEX yet).

6. Exploitation of the valency lexicon

The PDT-VALLEX and the VALLEX, two relatively independent applications of the principles of the shared framework, FGD, will prove the applicability of its valency theory.

In building the valency lexicon, stress is laid on comfortable and quick 'human readability', on easy orientation and intelligibility. On the one hand, such a format is necessary for an effective manual annotation, including discovery and correction of errors, on the other hand it should allow for a linguistically adequate representation and description of valency properties of verbs. Nevertheless, the main use of the lexicon is seen in automatic processing of Czech texts.

As for concrete applications of particular branches of the lexicon:

(i) The PDT-VALLEX serves for reaching the consistency of assigning the valency structure.

(ii) The contribution of the valency lexicon in an automatic syntactic analysis (parsing) is tested. Only a parser that can use valency information can make a difference between the structures of the sentences He began to love her. and He forced her to walk. Though these sentences have the same morphological annotation, they have different syntactic structures (different representation patterns on so called analytical level).

(iii) The valency lexicon is incorporated into an automatic system for creating an underlying representation of Czech sentences, a so called tectogrammatical parser. The main task of the respective module is to assign the functors to valency complementations and to add obligatory complementations that are not present in a morphemic realization of a sentence ('restoration of ellipses'), see (Hajič, Honetschlager, 2003).

(iv) Information on the valency properties of verbs stored in the lexicon is also exploited as source data for building the valency lexicon of nouns, namely as an input of an algorithm for conversion of verbal valency frames into valency frames of deverbal nouns. The algorithm can be applied on such nouns the valency properties of which undergo systemic changes, see (Řezníčková, 2003).

Acknowledgement

The research reported has been supported by the project of the Czech Ministry of Education LN00A063, Center for Computational Linguistics.

References:

Daneš, Fr., Hlavsa, Z. (1981) Větné vzorce v češtině. Academia, Praha.

Dorr, B.J. (2001) LCS Verb Database, Online Software Database of Conceptual Structures and Documentations, UCMP, http://www.umiacs.umd.edu/~bonnie/LCS_Database_Documentation.html

Fillmore, Ch. J. (1968) The Case for Case. In: Theory Bach, E., Harms, R. (eds.) Universals in Linguistic, New York, 1-90.

Fillmore, Ch. J. (1977) The Case for Case Reopened. In: Cole, P., Sadock, J. M. (eds.) Syntax and Semantics 8, New York-San Francisco-London, pp. 59-81.

Fillmore, Ch. J. (2002) FrameNet and the Linking between Semantic and Syntactic Relations. In: COLING 2002, Proceedings, pp. xxviii-xxxvi.

Hajič, J., Hladká, B., Pajas, P. (2001) The Prague Dependency Treebank: Annotation Structure and Support. In: Proceeding of the IRCS Workshop on Linguistic Databases, Univesity of Pennsylvania, Philadelphia, USA, pp. 105-114.

Hajič, J., Honetschlager, V. (2003) Annotation Lexicons: Using the Valency Lexicon for Tectogrammatical Annotation. PBML 79-80.

Hajičová, E., Hajič, J., Hladká, B., Holub, M., Pajas, P., Řezníčková, V., Sgall, P. (2001) The Current Statusof the Prague Dependency Treebank.. In: TSD2001 Proceedings LNAI 2166, Springer-Verlag, pp. 11-20.

Hajičová, E., Panevová, J., Sgall, P. (2002) A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank. UFAL/CKL Technical Report TR-2000-09, http://ckl.mff.cuni.cz/?a=techrep&m=publications

Hajičová, E., Kučerová, I.. (2002) Argument/Valency Structure in PropBank, LCS Database and Prague

Dependency Treebank: A Comparative Pilot Study. In: LREC 2002, Proceedings, pp. 846-851.

Kingsbury, P. Palmer, M. (2002) From TreeBank to PropBank. In: LREC2002, Proceedings, vol.VI., pp. 1989-1993.

Králíková, K. (1981) Reflexívnost sloves z hlediska automatické analýzy češtiny. Slovo a slovesnost 42, pp. 291-298.

Levin, B. (1993) English Verb Classes and Alternations: A Preliminary Investigation. Chicago: University of Chicago.

Ondruška, R., Panevová, J., Štěpánek, J. (2003) An exploitation of the Prague Dependency Treebank: a valency case. In: Simov, K., Osenova, P. (eds.) Proceedings of the Workshop SProLaC 2003, Lancaster University, pp. 69-77.

Mluvnice češtiny II., III. (1986, 1987) Academia, Praha.

Pala, K., Ševeček, P. (1997) Valence českých sloves. In: Sborník prací FFUB, pp. 41-54, Brno.

Pala, K., Ševeček, P. (1999) Česká lexikální databáze typu WordNet (v rámci projektu EuroWordNet-2). In: Sborník prací filosofické fakulty brněnské university, Brno, pp. 51-64.

Palmer, M., Rosenzweig, J., Cotton, S. (2001) Automatic Predicate Argument Analysis of the Penn TreeBank. In: HLT 2001, Proceedings, San Francisco: Morgan Kaufamm.

Panevová, J. (1974-75) On Verbal Frames in Functional Generative Description. Part I, PBML 22, pp. 3-40, Part II, PBML 23, pp. 17-52.

Panevová, J. (1994) Valency Frames and the Meaning of the Sentence. In: Ph. L. Luelsdorff (ed.) The Prague School of Structural and Functional Linguistics, Amsterdam-Philadelphia, John Benjamins, pp. 223-243.

Panevová, J. (1996) More remarks on control. Prague Linguistic Circle Papers 2, John Benjamins, pp. 101-120.

Panevová, J. (1998) Ještě k teorii valence. Slovo a slovesnost 59, pp. 1-14.

Panevová J. (1999) Česká reciproční zájmena a slovesná valence. Slovo a slovesnost 60, pp. 269-275.

Panevová, J. (2000) Poznámky k valenci podstatných jmen. Čeština - univerzália a specifika 2, Masarykova Univerzita, Brno, pp. 173-180.

Panevová, J. (2003) Some Issues of Syntax and Semantics of Verbal Modifications. In: Proceedings of MTT 2003, Paris. (in press)

Panevová, J., Řezníčková, V. (2001) K možnému pojetí všeobecnosti aktantu. In: Hladká, Z., Karlík, P. (eds.) Čeština – univerzália a specifika 3, Masarykova Univerzita, Brno.

Piťha, P. (1981) On the CyRyRase Frames of Nouns. Prague Studies in Mathematical Linguistics 7, Academia, Prague, pp. 215-224.

Piťha, P. (1982) K otázce valence u adjektiv. Slovo a slovesnost 43, pp.113-118.

Řezníčková, V (2003) Czech Deverbal Nouns: Issues of Their Valency in Linear and Dependency Corpora. In: Simov, K., Osenova, P. (eds.) Proceedings of the Workshop SProLaC 2003, Corpus Linguistics 2003 conference, Lancaster, pp. 88-97.

Sarkar, A., Zeman, D. (2000) Learning Verb Subcategorization from Corpora: Counting Frame Subsets. In: LREC2000 Proceedings, vol.I, pp. 227-233.

Skoumalová, H. (2001) Czech syntactic lexicon. PhD thesis, Charles University, Faculty of Arts, Prague.

Slovesa pro praxi. (SPP) Svozilová, N., Prouzová, H., Jirsová, A, Academia, Praha, 1997.

Slovník české frazeologie a idiomatiky (SČFI) Academia, Praha, 1983.

Slovník českých synonym (SČS) Pala, K., Všiansky, J., Lidové noviny. Praha 1994.

Slovník spisovné češtiny pro školu a veřejnost (SSČ) Academia, Praha, 1978.

Slovník spisovného jazyka českého (SSJČ) Praha, 1964.

Sgall, P., Hajičová, E., Panevová, J. (1986) The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Dordrecht: Reidel, Prague: Academia.

Straňáková, M. (2001) Homonymie předložkových skupin a možnost jejího automatického zpracování. Disertační práce MFF UK.

Straňáková-Lopatková, M, Žabokrtský, Z. (2002) Valency Dictionary of Czech Verbs: Complex Tectogrammatical Annotation. In: LREC2002, Proceedings, vol.III., pp. 949-956.

Tesnière, L. (1959) Élements de syntaxe structurale. Paříž.